# CS 242: Information Retrieval & Web Search

# Build a Search Engine

Team Member(s):

James Nacino

Scott Michael

## Project Composition

I would like to crawl web pages on Yelp.com. Yelp has many different types of business listings ranging from restaurants, dental offices, and entertainment. Specifically, I would like to crawl restaurant web pages on Yelp. I am not sure yet if I want to focus on restaurants that sell specific cuisine items such as Italian, Mexican, Asian, or American. However, the main goal is to gather as many reviews from each restaurant using Jsoup. I will store the data that I gathered in a database. Most likely a RDBMS, or NoSQL database.

Along with gathering the reviews, I will make sure to store the associated restaurant name and the location of the restaurant. As a secondary goal I will try and store some images of foods or maybe a picture that restaurant. Using Apache Lucene, I will be able to implement some ranking model on the documents. The documents just to be clear, would be the restaurants with their associated location and reviews.

I would want to be able to be able to use two different textboxes when inputting their query. The goal of my query will have the user to be able to fill out the “location” and “search” textboxes. The “location” textbox will be responsible for filtering restaurants within a specific city and the “search” textbox will be used by the ranking algorithm to find the most associated restaurants that have the query in the review. I will make the “location” textbox not mandatory to fill out - if they leave this textbox blank then we will return the most associated restaurants just based on the “search” textbox which ranks restaurants based on the key words that the user inputted.

The final solution will look something like below on the next page:

